FUNDAMENTALS OF MULTIMEDIA COMPUTING

Chapter 2: Elements of Multimedia Computing

Authors:
Gerald Friedland
and
Ramesh Jain

DRAFT FOR COMMENTS

Chapter 2
Elements of Multimedia Computing

Introduction

Multimedia is closely related to how humans experience the world. In this chapter first we introduce the role of different sensory signals in human perception for understanding environments to function in it and for communicating and sharing experiences. A very important lesson for multimedia technologists is that each sense provides only partial information about the world. Data and information from different sensors must be combined with other senses and prior knowledge to understand the world. One sense alone, even the most powerful sense of vision, is not enough to understand the world. In multimedia computing also, different sensory media should be combined with other knowledge sources to interpret the situation. *Multimedia computing and communication is fundamentally about combining information from multiple sources in the context of the problem being solved.* This is what distinguishes multimedia from several other disciplines, including computer vision and audio processing, where focus is on analyzing one medium to extract as much information as could be extracted from it.

In multimedia systems, different types of data stream simultaneously exist and the system must process them not as separate streams but as one correlated set of streams that represent information and knowledge of interest for solving a problem. *The challenge for a multimedia system is to discover correlations that exist in this set of multimedia data and combine partial information from disparate sources to recover the holistic information in a given context.*

Experience and Information

We experience our physical environment through our natural senses of sight, sound, touch, taste, and smell. Every human child starts building the models of different objects and events in the world through learning process from very early part of life using all senses. Once these models are in place, our senses let us experience and function in the physical and social worlds and refine, enhance and even develop new models. These models are fundamental to recognition of objects and events in our world. Model of an object or event helps us in abstracting all sensory information into a simple symbol. The process of assigning a symbol to represent an object or event and then building more complex objects and events using these is at the heart of a field called *semantics*. Semantics is the study of meaning. Semantics is an important area of study in linguistics. Semantics is related to the study of meaning of words, phrases, sentences, and other larger units of text. Semantics is a rigorously studied field [8]. In our discussions in this book, we will not address detailed theory of semantics but we will consider the basic aspects as needed. For our purpose, we will just consider semantics to be the study of meaning associated with symbols. These symbols could be simple atomic symbols or could be composite symbols built by combining multiple atomic and/or composite

symbols. Since our concern is with multimedia signals, these symbols could represent some units in different components such as audio and visual data or could represent entities as combination of different media thus resulting in symbols in multimedia.

Webster's dictionary defines experience as the "direct observation of or participation in events as a basis of knowledge." We experience the world we live in. The basis of all our interactions is our experience of the world. We learn about the world and accumulate and aggregate our experiences in the form of knowledge. Scientists among us *experi*ment to test their knowledge and to gain new knowledge. Scientific process relies on experiments to study a hypothesis under different conditions to evaluate its validity. Experimental aspects of a science are fundamental to its progress. Final evaluation of experiments is by humans using their sensory processes.

Communication is the process of sharing experiences with others. The history of civilization follows the development of our understanding of experiences and how to share them with our fellow humans even in other parts of the world immediately, as well as with those who will follow in future generations. It is interesting to see how many influential innovations and inventions in human history are related to communication and sharing of experiences with people who may be spatially and temporally separated. This process started with the development of languages and has lead to the innovations resulting in the World Wide Web.

Information is an "efficient but abstract communication of experience." We gain knowledge through the set of experiences that make up our lives and communicate information about those experiences. Because we don't communicate the experiences themselves, we lose a vital element of the act of experiencing in the translation to information. Many of us can't know, for example, what it's like to hit the game's winning run, to surf the perfect wave, or to fight a war. We can read about the events, of course, and we can view pictures and videos. But we aren't present in these, so our experience and hence the knowledge of these events is incomplete.

In the communication process, one of the most important elements is to develop a dictionary. A dictionary is an exhaustive collection of a selection of the words of a language. It contains information about their meanings, pronunciations, etymologies, and inflected forms, in either the same or another language. Thus, a dictionary is a shared and agreed collection of symbols (words) and what these symbols mean. Each language is based on a dictionary containing these symbols and rules, the grammar of the language, to use them. Without a dictionary communication may not happen. Just imagine situation when a person speaking English is talking to a person speaking Chinese. Each person is using a dictionary but these are two different dictionaries. Communication requires a shared dictionary. Dictionaries are commonly used in the context of languages and use words as the basic symbols as carrier of meaning. In computer science, dictionaries are extended to use list of codes, terms, and keys for use by computer programs.

For regular language dictionaries, many electronic dictionaries now include audio pronunciation for words. This is first step in including audio in dictionaries. Multimedia dictionaries are likely to evolve and contain features in one or more mediums to form alphabets for that medium and develop approaches to construct multimedia worlds that will represent concepts in real world.

In multimedia, the basic carrier of meaning or symbols are not just traditional words as used in speech and text, but some units, similar to alphabet used in text and phonemes used in speech, in a particular medium. Let's consider visual information. Consider a very simple common task: Given an image of an object, name this object and list all objects that look like it and are related to it. Try to extend this to all detailed functions that you commonly see in a dictionary. Visual Dictionaries are being developed for different applications and for different kinds of visual information, ranging from shapes of curves to complex objects [7]. In these dictionaries, usually visual shapes or objects are stored and all their details are given in multiple languages. It is likely that these dictionaries will play increasingly important role in understanding of visual information and applications of emerging technology that will increasingly utilize cameras as information devices.

Perception

Perception is the process of understanding sensory signals for recovering information [6]. Perceptual processes have been analyzed with the goal to understand them for very long time. With arrival of computing, it attracted more attention from psychologists and researchers in artificial intelligence in the hope of developing machines for automatic perception. The understanding of perceptual processes has remained a difficult problem and is a very active research area in many disciplines including psychology, neuro-physics, and computer science. Understanding of sensory information is a very important step in many multimedia systems. We will study important perceptual processes in audio and visual processing in following chapters. Here we present some general aspects of perceptual processes.

A perception system takes sensory signals as input and generates the information that is relevant in its application context as output. This process is based on two important sources: signals and relevant knowledge sources. Figure 1 shows the role of existing knowledge in perception. The output of the system is clearly the combination of the input sensory signal as well as the knowledge available to the system. Without any knowledge, the system can not produce any information, and without the signal the system can only hallucinate. Perception sometimes is considered as a controlled hallucination process [3] where based on the signal the system starts hallucinating and creates multiple hypotheses then uses the signal to find the best support for its hypotheses and recovers information from signal.

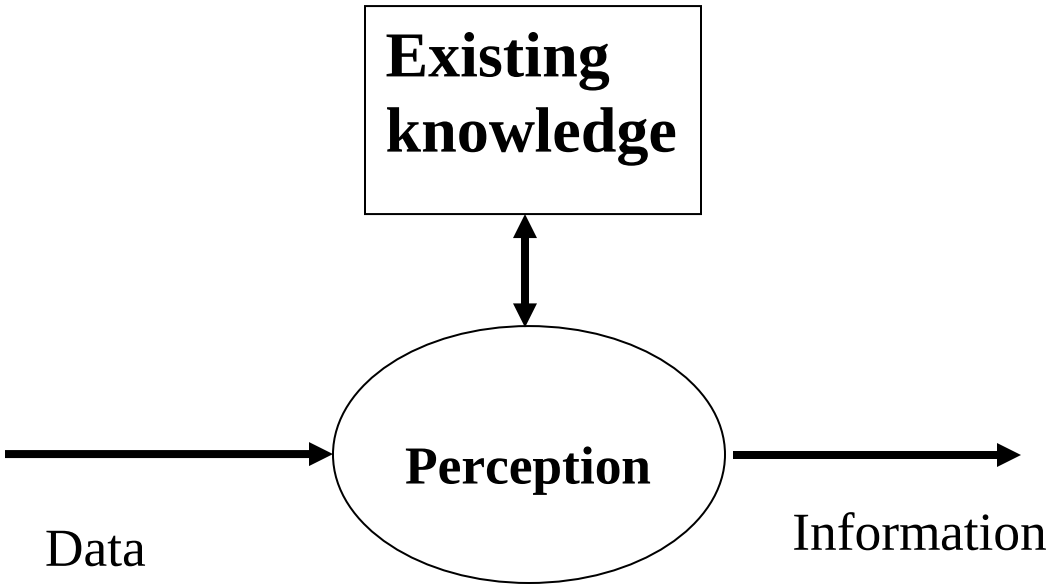

Figure 1. Perception is the process that recovers meaningful information from sensory data using existing knowledge.

A **B**

Figure 2: (A) A Dalmatian dog sniffing around. (B) Unstable perception: two faces or vase.

The role of knowledge, in the form of models of objects and events, is not immediately obvious. Some examples may make it very clear, however. We always use the term *recognition* for finding objects in signals, such as images. This term implies that we try to *re cognize* objects – meaning we know about the object or in other words have models of objects. Without models, there is no recognition. The models could be in many

different forms ranging from very concrete and detailed model to very abstract models. To show the importance of models, we show two very commonly seen pictures in Figure 2. In Fig 2 a, there is a Dalmatian dog sniffing around. If you don't know how a Dalmatian dog looks, you will see only random blobs in this picture, but if you know Dalmatian dog, you will clearly see it. The Fig 2 b shows the picture which has two interpretations: you can either fit model of human faces to it and see two faces or see a vase in it. This shows that your perception system comes up with two hypotheses and admits both as viable, but only from a slightly different gaze point.

Perceptual Cycle

In all our activities, we use our five senses, brain, and memory to understand our environment, to operate in this environment, to communicate about the environment and finally to build and update our knowledge repositories for efficient and effective use of what we learn. How we use these senses and how we convert this sensory data to information and knowledge has been a source of intrigue to thinkers for almost all known history. Here we present some ideas to provide historical context and perspectives on the evolution of this understanding.

Even ancient philosophers [5, 13] believed that:

- Understanding of world is indirectly derived using sensors.
- The fidelity of the model of the world depends on how well a person understands the world.
- People achieve different 'levels' of understanding in terms of their own knowledge structures.
- Nirvana is the highest stage of understanding.

These observations show deep insights about 2000 years ago. Thinkers even during that time clearly recognized that data from all sensors must be assimilated using existing knowledge to form an understanding of the environment. It was also recognized that sensors help us understand the world at different levels of understanding. One evolves to the highest level of understanding by refining their knowledge structures. Similar ideas and models are discussed by modern philosophers [10] in theory of objective reality.

To formulate the problem from computational perspective, we consider perceptual cycle introduced by Ulric Neisser [2,8] in 1976 to model how people perceive the world. According to this model, a perceiver builds a model of the world by acquiring specific signals and information to accomplish a task in a natural environment. The perceiver continuously builds and refines a schema that is based on the signals he received so far. This schema represents the world as the perceiver sees it at that instant. The perceiver then decides to get more information to refine the schema for accomplishing the task that he has in mind. This sets up the perceptual cycle in Figure 3 below. The basic idea behind the perceptual cycle is that an agent is continuously interacting with the environment using its sensory mechanisms to build the model of the environment that

will be useful in its task. At any given time instant it has a model of the environment, called schema that is constructed using all the data and information received until that point. The system then decides what more is required to complete the task and how that information could be acquired. Based on this the agent collects more information using appropriate sensors.

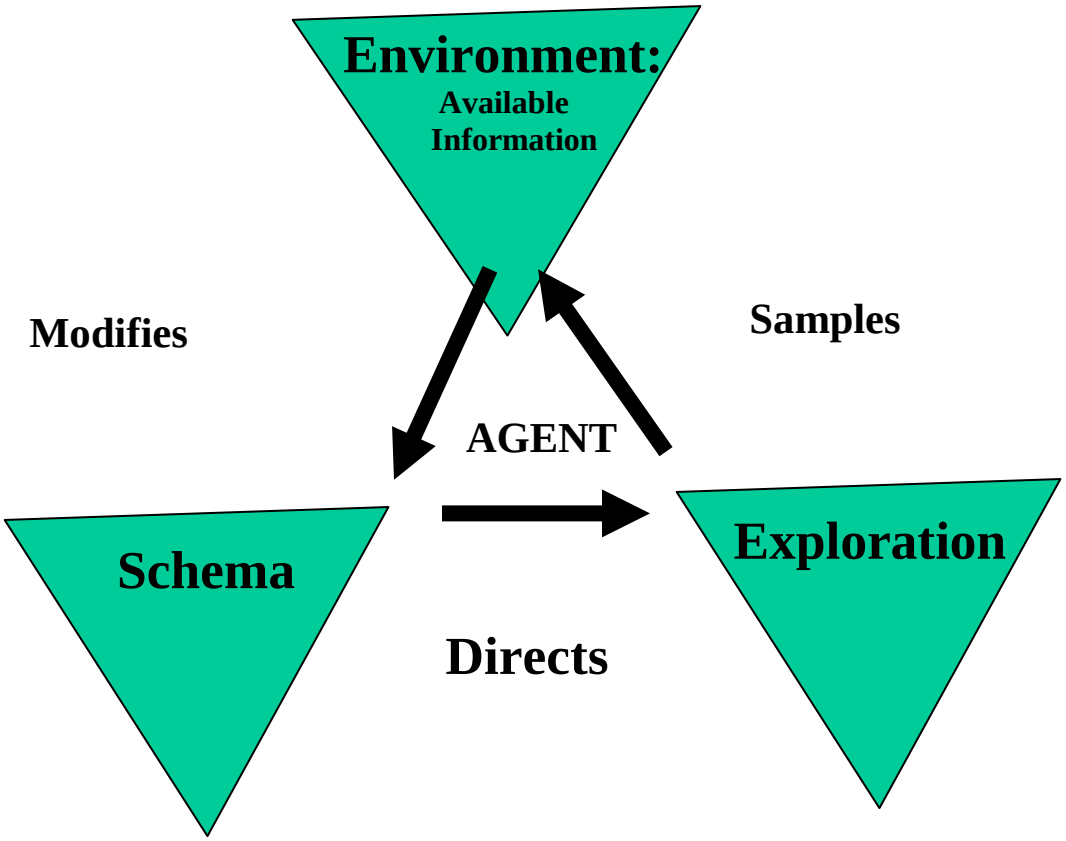

Figure 3. Neisser's perceptual cycle: The perceiver gets signals from the environment, interprets them using the current schema, uses the results to modify the schema, uses the schema to decide to get more information, and continues the cycle until the task is done.

The perceptual cycle model has conceptual similarity to recursive filtering techniques commonly used to estimate the state of a linear dynamic system using observers (sensors) that may provide noisy measurements. The state of the system is represented mathematically as a vector. The state vector represents the values of the parameters that are used to characterize the linear dynamic system. In system theory [1], these vectors

represent the system so that correct amount of control inputs could be applied to the system to bring it into the desired state. In perceptual cycle, the schema represents the essential parameters that are required to solve the given task. Based on the current schema as compared to the final, or desirable, schema the agent must decide its action.

As mentioned, however, the perceptual cycle is dealing with perception that is not a linear dynamic system. This can not be easily modeled using the current tools of the system theory. Some powerful estimators, such as Kalman filters [1], have already been used in computer vision to model aspects of human perception. As progress in technology takes place, it is expected that more formal tools will be developed to represent and construct schema using multimedia data.

Multimedia Systems

Consider a computing system equipped with multiple sensors working in a physical environment. The system continuously gets information about the environment from multiple sensors and must process all these in the context of its application. Obviously the applications could range from just identifying an object to autonomously functioning in a complex dynamic environment. Here we consider a general situation without any specific application. We also consider that for a computing system, the types of data sources are not limited just to the sensory modes that we humans can process. As is well known, different species have different sensory capabilities. Multimedia computing systems could be endowed with sensing capabilities of various types.

Let us assume that we are given $S_1, \ldots S_n$ data streams. These data streams have K types of data in the form of image sequence, audio stream, motion detector, annotations, symbolic streams, and any other type that may be relevant and available. We assume that these streams can be synchronized both in time and space resulting in these data streams represented in a common temporal and spatial coordinate system rather than in the coordinate system of each sensor. Further, we assume that metadata $M_1, \ldots, M_n$ for each stream is available from the original sources that helps us in interpreting the data stream in the context of the world. This meta data may include things like location and type of the sensor, viewpoint, angles, camera calibration parameters or any other similar parameters relevant to the data stream. Data stream is usually not directly very useful in the interpretation of the data in relation to the environment. In most cases, some feature detectors must be applied to each data stream to obtain features that are relevant in the current environment. Le us represent, features stream F_{ij} as the j^{th} feature stream in S_i.

Given the above data environment, the most fundamental problem that multimedia computing systems must solve is to extract relevant information about the task at hand using data from these disparate sources. There are many challenging problems, including the following that are directly relevant to the main theme that we will address in this book:

- How do we combine these data streams to obtain the information that is essential for solving the problem at hand?
- How do we represent this data in the most compact form for communication and storage?
- How do we present this volume of data to a user in her computing environment to communicate intended information?
- What are the system issues that must be solved to deal with these disparate types of data and how they are handled by the system?

Before we address most specific concepts and techniques related to solving above problems in the rest of this book, some concepts that form the basic fabric of multimedia systems are discussed in the rest of this chapter.

Semantic Gap

Computing systems represent data in terms of bits and bytes and build from these more sophisticated representations such as lists, images, audio, and video. All these representations are fundamentally a collection of bits that programmers use to define abstractions to build their applications. On the other hand, the users of these systems are people who define their applications in terms of objects and events and build complex concepts based on abstractions that start with objects and events. There is a fundamental gap between the abstractions defined in computing systems and those used by the users of these systems. This situation is shown in Figure 4. This gap is defined in [15] as:
"*The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation. A linguistic description is almost always contextual, whereas an image may live by itself.*"

In current computers, we must build abstractions starting with bits, the most fundamental representation unit of data, and defining concepts that may be needed in specific applications. It is easy to build these concepts by defining various structures and naming and using them as a programmer may want to. We are all familiar with, and will discuss more in following chapters, concepts such as images, video, and audio signals as they are represented in computers. Human beings, currently usually ultimate users of the computing systems, usually do not think in these structures, however. Humans usually think in terms of objects and events [11] and build complex concepts based on complex, often ill-defined and uncertain, relationships among them. As shown in Figure 4, there is a gap between the abstractions such as images and video as defined in computers and objects and events as used by people in their mind. This gap is what is called the semantic gap.

The main reason for the semantic gap, which may even exist among two persons, is that the physical world contains objects and events and people build the models of these objects and events in their mind based on the data that they receive from their sensors. People learn to abstract the data received from their sensors in terms of objects and events while combining this data naturally from all sources including all sensory organs,

context, and memory. On the other hand, we try to define the models of these in computers using the data that is represented fundamentally in units of bits. The abstractions that are built in computers are based on what could be built in computers using bits. Fundamentally, in computing we define things based on what could be computed, while as humans we learn to abstract what is needed to live.

Many concepts and techniques developed in multimedia computing are related to bridging the semantic gap. Starting from signal analysis approaches used in audio processing and computer vision to indexing for multimedia information retrieval, many concepts and techniques in multimedia address the problem of bridging the semantic gap. In fact in all those cases where human beings are integral part of a computing system, the semantic gap must be bridged. In many mature fields, this is bridged either by developing concepts in the field that bring data and humans conceptually close or by developing interfaces that rely on human intelligence to bridge the gap.

Consider common search systems that appear to work so well. A close look at search system's behavior shows that when searching for keywords that can be matched as character strings, it is easy to get good results. When you are searching for something that will not be satisfactory only based on string matching, but that requires some interpretation of either data or your intentions, search systems perform poorly. Most research in improving relevance of results in search engines is trying to bridge semantic gap. It is concerned with how to interpret data and how to detect a user's intentions based on contextual information.

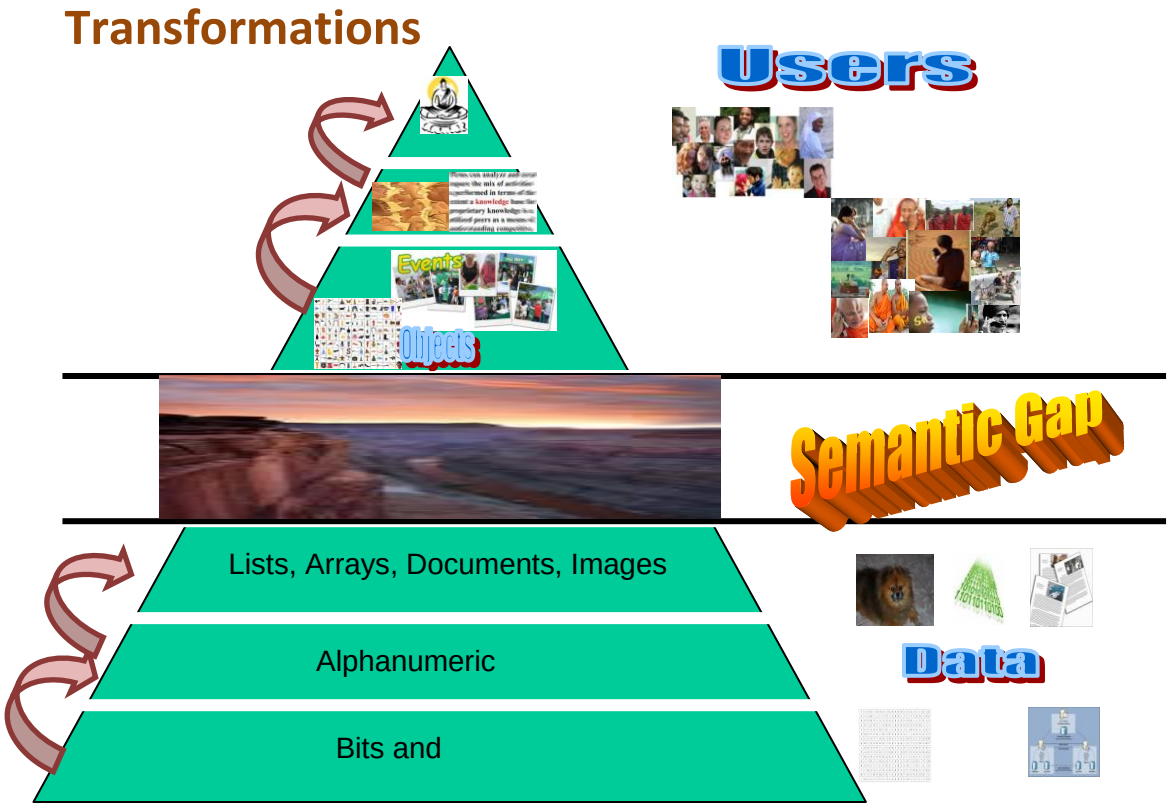

Figure 4: Semantic Gap. There is a big gap in how computers represent data like images in bits and bytes and how people think about images as collection of objects or events.

Figure 5: This photo was taken at **Zhujiajiao, Shanghai , China.**

Context and Content

Content and context are two very commonly used terms in multimedia processing and understanding. There is no rigorous formal definition of content or context, though they are used extensively by practitioners and researchers. It is important to understand what they mean and how they are related to understand and develop efficient and effective multimedia systems.

Given a file containing a picture, or an audio, or an article (text); content is what is contained in the file. Let's consider a picture file – say a photo shown in figure 5. This is a photo containing 4224 X 3168 pixels and each pixel is a color pixel. This photo contains more than 13 Million points (Pixels) each with three color values associated with it. In the most basic form the content of the file or, as commonly used, photo are these 13 M pixel colors in the spatial configuration as represented by the location of these pixels.

This photo was taken at Time (11:31 AM on Sept 13), Location (Zhujiajiao, Shanghai: Latitude 31 deg 6' 36.34" N, Longitude 121 deg 2' 59.22" E) and using a Nikon Coolpix P6000 camera. For this particular photo, no flash was used, and the focal length (6.7 mm), ISO (64), and aperture (f/4.7) values of the camera are known. All this (and much more) information is captured by the camera and is stored in the picture file along with the above pixel data using a popular EXIF (Exchangeable image file format) standard [4]. This data, commonly called metadata, represents the context in which the data in the photo was acquired. Context is defined as the interrelated conditions in which some data (the content) is acquired. As seen above, some context parameters are stored by the camera using EXIF standards for the photo. EXIF standard is used almost by all digital camera manufacturers to store this kind of data with all digital photos.

Some other context parameters may help in understanding of data. For example, in the context of the above picture, EXIF tells the model of the camera but it will be very helpful if the owner of the camera is known and profile and calendar information about the owner is also available. In many cases based on this information it may be possible to understand what the objects are and more importantly, who the person in the picture is.

Multimedia research and techniques developed were concerned with only the content of the data in early days. Increasingly the importance of context is becoming clear. Recently [14] researchers are emphasizing that content and context should be combined and should be viewed as all information that must be used for understanding multimedia.

Meta Data

Metadata literally means 'data about data'. Given some data that represents an audio, photo, or video; the metadata for it will describe different characteristics of this data. Some obvious metadata is name, size, date of creation, and type of the file. In addition to these, one can include any other information that is considered useful for understanding,

storing, transmitting, presenting, or any other operation on the file containing the data. Metadata itself is data of some particular type; it is metadata in this particular context because it is used to qualify or help some other data.

Since understanding techniques for text, audio, images, and other sensory data have not matured enough to correctly understand elements in data, metadata has gained in popularity particularly with the growth of the Web. XML was designed to be a mechanism to transport and store data. It accomplishes that by defining a language to describe what data is. Tags used in XML are data about data. Thus, XML is a language to associate metadata with the data explicitly so it could be read by any program. This helps in not only transporting and storing, but analyzing and understanding data.

In multimedia computing, use of metadata is increasing rapidly. Many approaches based on XML and tags are becoming commonplace even in audio, images, and video. It is expected that techniques to represent metadata as intimately as in text will evolve in this area. Use of EXIF with all stored digital images is a clear example of this trend in this field.

Objects and Events

In understanding data of any type, one tries to find which aspect of the world the data represents. As discussed above, perceptual processes depend on prior knowledge about the world we live in to analyze the signals. An important issue is how to represent the world.

Many researchers believed that the world can be represented using objects. This view believed that the world could be considered as a collection of objects. This view is challenged by many modern, and not so modern, thinkers [11]. According to their views, events play equally important role. Events represent change in relationships among objects. And these changes are fundamental to understanding the current state of the world. According to emerging views, to model dynamic world both objects and events must be used. In a sense, objects are good in capturing static component of the world, while events complement that by capturing dynamic situations.

In computer science, object oriented thinking has been used in many fields and their applications. Object oriented approaches have revolutionized many areas of computer science because of the high level abstractions it offers for design, programming, and even some interactions.

Multimedia brings some new challenges to computer science, however. Multimedia, particularly audio and video, are fundamentally dynamic in nature. They capture signals that represent some attributes of the world as a function of time. In fact in many applications, even those sensors that capture static characteristics of the world, such as temperature at a point in space and time, are used to detect changes in those characteristics as function of time. A sensor is almost always placed in an environment

where some event needs to be detected and the sensor measures some physical attribute that helps in detecting the event.

Many fields in computing have used the concept of event in designing systems. This concept has been used very differently in different fields, however. With increasing use of multimedia in computing, it is likely that a unified approach for event-based thinking will evolve [16,17].

It must be emphasized that for modeling real world using powerful computational approaches, it is essential that both objects and events be used. Objects and events complement each other. Objects in computing capture attribute and functions of physical objects and other related concepts and events represent relationships and changes among those relationships in space and time.

REFERENCES

1. D. E. Catlin, “Estimation, Control, and the Discrete Kalman Filter”, Springer Verlag, 1988.
2. Igor A. Chimir, Waheeb A. Abu-Dawwas, Mark A. Horney, “Neisser's cycle of perception: formal representation and practical implementation”, Journal of Computer Science (Special Issue): 106-111, 2005
3. Clowes, M. B. (1971), "On seeing things," *Artificial Intelligence*, Vol 2, No. 1, pp 79--112.
4. http://www.exif.org/
5. Hermann Kuhn, *The Key To The Center Of The Universe*, Crosswind Publishing, 2001.
6. James J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, 1987.
7. Jonathan Metcalf, *Five Language Visual Dictionary*, 2003.
8. Neisser, U (1976) Cognition and reality: principles and implications of cognitive psychology WH Freeman
9. Jaroslav Peregrin *Meaning: The Dynamic Turn. Current Research in the Semantics/Pragmatics Interface*. London: Elsevier, 2003.
10. Karl Popper, “Three World”, The Tanner Lecture on Human Values Delivered by Karl Popper at The University of Michigan on April 7, 1978.
11. A Quinton, “Objects and events”, Mind, pp. 197-214, 1979.
12. Irvin Rock, *The Logic of Perception*, MIT Press, 1985.
13. L. M. Singavi, *That Which Is: Tattvartha Sutra (Sacred Literature),* AltaMira Press, 1998.
14. P. Sinha and R. Jain, “Semantics in Digital Photos: A Contenxtual Analysis”, Int. Journal of Semantic Computing, Vol2, No. 3. pp 311-325, 2008.
15. Arnold Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain “Image Databases at the end of the Early Years” IEEE Transactions on Pattern Analysis and Machine Intelligence 23(1), January 2001.
16. A. Scherp and R. Jain, “An Eco System for Semantics,” in IEEE Multimedia, June 2009.
17. G. Utz Westermann and Ramesh Jain,” Towards a Common Event Model for Multimedia Applications”, in IEEE Multimedia, January 2007.

Exercises

1. What is semantic gap? Why does semantic gap become a serious problem in perceptual systems?

2. A digital camera collects a lot of meta data related to camera parameters, including its location, and stores that with the intensity vales at every pixel. How can the meta data be used? Can this meta data help in reducing the semantic gap?

3. What is EXIF? Where can EXIF be useful?

4. How is text related to audio?

5. Which is easier to analyze, speech or text? Why?

6. What is the role of knowledge in perception system? List at least 3 knowledge sources that could be used in understanding images?

7. What is a model as used in perception systems? Can you develop a recognition system without using a model?

8. What is a perceptual cycle? How is it related to estimation theory?

9. Multiple sensors are usually used to capture attributes of real world. Since the sensors have different coverage in space and have different temporal characteristics, how can one combine the data obtained from these sensors?

10. What is a feature in a sensor data stream? What role does it play in analysis of data and correlating different data sources?